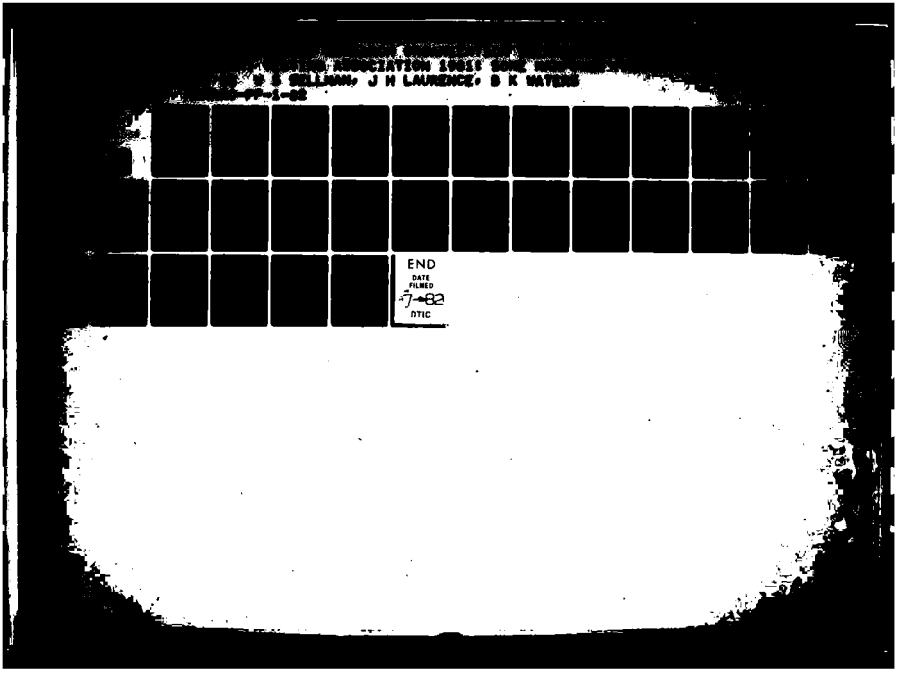
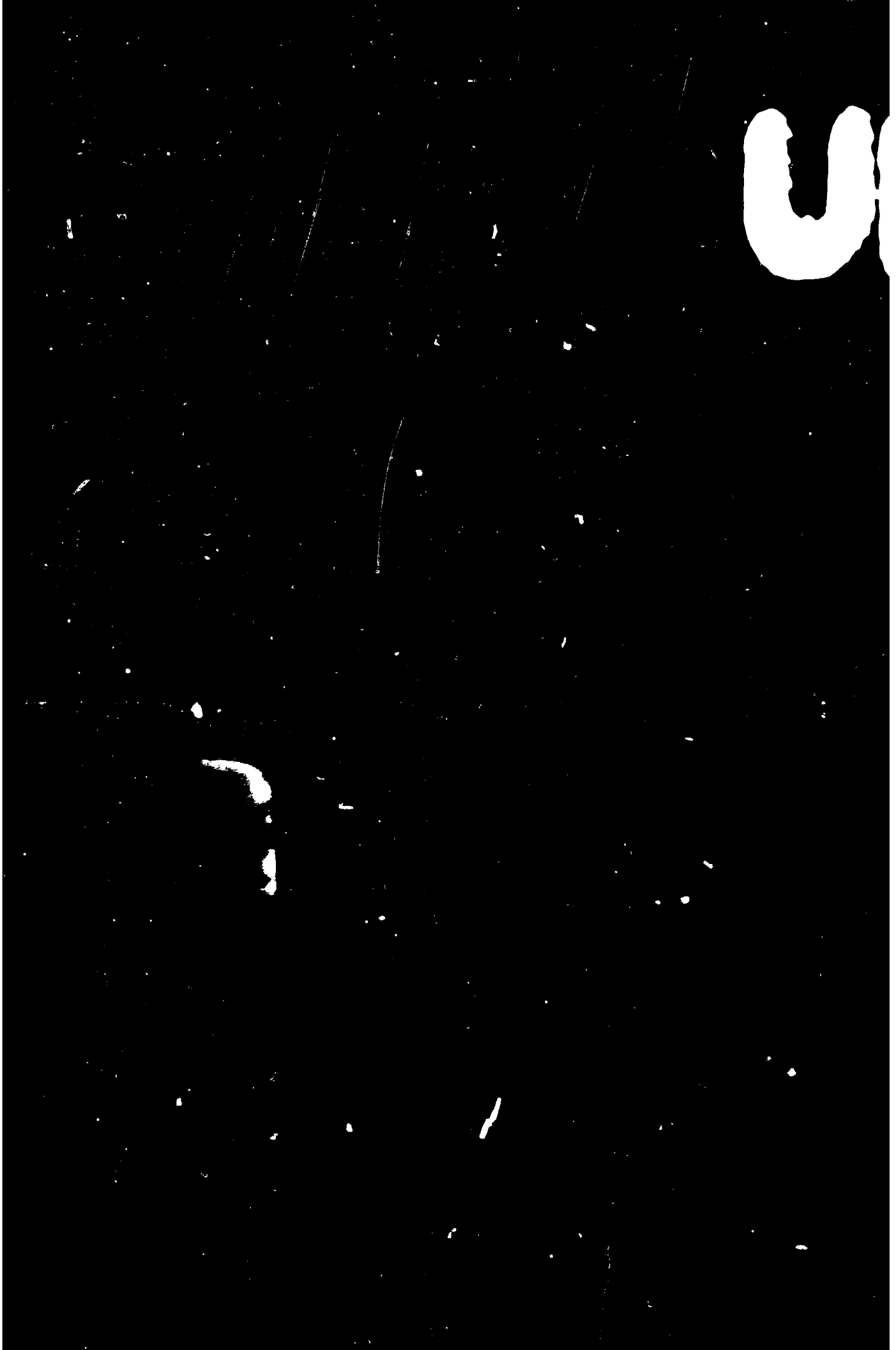
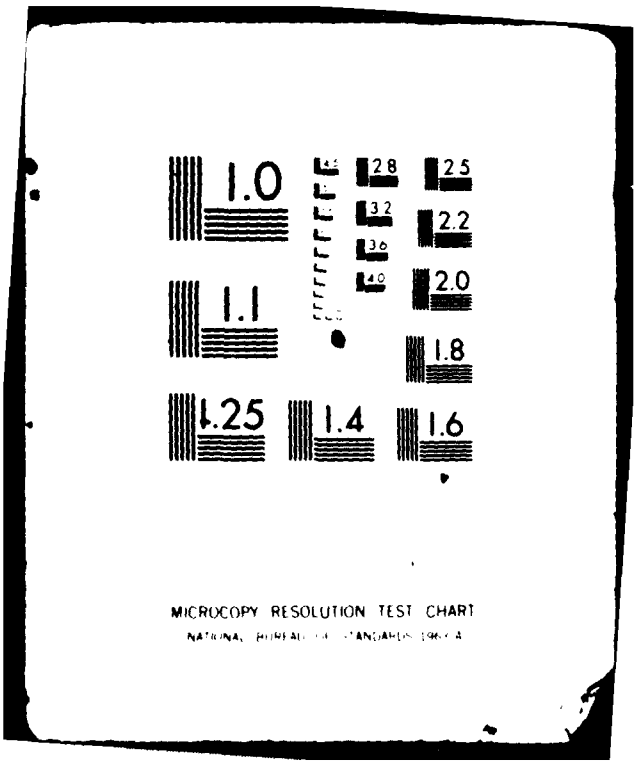

END
DATE
FILMED
7-82
DTIC




1.0
1.1
1.25
1.4
1.6
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.8
MICROCOPY RESOLUTION TEST CHART

Professional
Paper
1-82

HumRRO-PP-1-82

HumRRO

Military Testing Association 1981: Some Manpower Presentations

Wayne S. Sellman
Office of the Assistant Secretary of Defense,
(MRA&L)

Janice H. Laurence
Brian K. Waters
Mark J. Eitelberg
Gus C. Lee
Human Resources Research Organization

Four papers presented at the
23rd Conference of the Military
Testing Association
Arlington, Virginia October 1981

HUMAN RESOURCES RESEARCH ORGANIZATION
300 North Washington Street • Alexandria, Virginia 22314

March 1982

(12)

PREFATORY NOTE

This paper is based on four presentations given at the 23rd Conference of the Military Testing Association by research scientists in HumRRO's Manpower Program Analysis Division.

The first three papers, "Aptitude Testing in DoD and the Profile of American Youth Study," "Military and Civilian Test Score Trends (1950-1980)," and "Subpopulation Analyses of Current Youth Aptitudes," were presented at the symposium entitled, Profiling the Aptitudes of the Current Youth Population.

The fourth paper, "Legal and Political Considerations in Large-Scale Adaptive Testing," was presented at the symposium on Psychometric Considerations for Adaptive Testing Systems: Administration and Validation.

The Military Testing Association conference was held at Arlington, Virginia, October 25-30, 1981, and was coordinated by the U. S. Army Research Institute for the Behavioral and Social Sciences.

APTITUDE TESTING IN DoD AND THE PROFILE OF AMERICAN YOUTH STUDY

Wayne S. Sellman
Office of the Assistant Secretary of Defense
(Manpower, Reserve Affairs, and Logistics)

Janice H. Laurence
Human Resources Research Organization

INTRODUCTION

Discussions of present or future military manpower procurement policies consider the way in which individuals are selected for service, assigned to military jobs, and trained to perform those jobs. Philosophically, there is consensus that enlistment standards are essential for manning an effective military. Beyond that broad agreement, the type and kind of enlistment standards (i.e., medical, moral, educational, and aptitude) are topics for ideological, legal and scientific debate.

The Armed Services have devoted considerable effort to develop reliable and valid methods for assessing persons prior to their entering military service. One focus of these efforts has been on the development of tests which measure the aptitudes of individuals. Aptitudes have historically been defined as measures of trainability for the various military jobs.

Aptitude levels within the military have been referenced statistically to the extensive testing of adult males that took place during World War II. This World War II "reference population" has been the baseline for comparing aptitudes of military examinees and recruits across time. Recently, questions have been raised concerning the appropriateness of retaining the World War II reference population as the sole basis for today's military personnel decisions. Accordingly, it was decided that the contemporary youth population should be examined to facilitate the Department of Defense's understanding of the quality and representativeness of its new enlistees.

An aptitude profile of current youth is important for recruiting, and evaluating recruiting results. The Department of Defense (DoD) should be able to compare the characteristics of today's population with DoD requirements for military manpower. Information is also needed for mobilization planning. If a national emergency required the resumption of conscription, DoD must be able to establish entrance standards compatible with available manpower resources that meet the personnel needs of the Services. Decisions on who should be drafted or permitted to volunteer require an accurate knowledge of the aptitudes of contemporary youth.

The 1980 Profile of American Youth. The Profile of American Youth was designed to assess the vocational aptitudes of young people, ages 16 to 23, and, at the same time, to develop a new reference population against which scores on the DoD enlistment test can be interpreted. To achieve these

goals, DoD contracted with the National Opinion Research Center (NORC) of the University of Chicago to administer the enlistment test to a nationally representative sample of about 12,000 young men and women.

Beyond its value to military manpower planning, aptitude profiles from a national sample of young people are a significant contribution to scientific research. Such aptitude profiles have not been previously available due to the difficulty and expense in obtaining representative data.

APTITUDE TESTING IN DOD

The Armed Services Vocational Aptitude Battery. The test used in the 1980 aptitude profile study was the Armed Services Vocational Aptitude Battery (ASVAB). ASVAB was introduced on January 1, 1976 as the single DoD test to replace the various aptitude test batteries then in use by each Service. Replacement forms were subsequently implemented on October 1, 1980. A 1980 version (Form 8A) of ASVAB was administered in this study.

ASVAB scores serve two important purposes in the enlistment process. First, they help determine eligibility for enlistment. Second, they are used to establish qualifications for assignment to specific military jobs.

The ASVAB consists of the following 10 subjects: arithmetic reasoning, numerical operations, paragraph comprehension, word knowledge, coding speed, general science, mathematics knowledge, electronics information, mechanical comprehension, and automotive-shop information. These subtests are included because research and experience have shown them to be valid predictors of success in military training.

The scores of four subtests (word knowledge, paragraph comprehension, arithmetic reasoning and numerical operations) are combined to produce an Armed Forces Qualification Test (AFQT) score. The AFOT score, supplemented by the scores on various aptitude composites, are used in conjunction with educational, medical, and moral standards to determine applicant enlistment eligibility. The scores on the aptitude composites also determine eligibility to enter specific military skills.

The Services combine a variety of subtests to form aptitude composites. Table 1 shows the subtests that comprise two selected composites.

Table 1

Selected Aptitude Composites and Their Component ASVAB Subtests

Selected Composites	ASVAB Subtests
Administrative	Paragraph Comprehension
	Word Knowledge
	Numerical Operations
	Coding Speed
Electronics	Electronics Information
	General Science
	Arithmetic Reasoning
	Mathematics Knowledge

The Armed Forces Qualification Test. During the early years (1940-1942) of World War II, men were accepted for service if they had completed the fourth grade or were able to pass literacy screening tests; in later years (1943-1945), minimal literacy was no longer required for induction (Ginzberg, Anderson, Ginzburg & Herma, 1959). After service entry, the primary test instrument for job assignment purposes was the Army General Classification Test (AGCT). A test of general trainability, the AGCT was composed of questions which measured verbal, arithmetic, and spatial abilities. After World War II, it was used by the Army for enlistment screening. Modeled after the AGCT, the Armed Forces Qualification Test (AFQT) was introduced in 1950 to determine the eligibility or draftees and volunteers to enter any of the Services (Uhlaner & Bolanovich, 1952).

To minimize test compromise and to update test language and content, the AFQT has been revised periodically. Until 1973, each new AFQT was calibrated back to the AGCT so that successive AFQT scores would have a constant meaning in terms of the level of trainability. In 1972, the use of a common AFQT was discontinued. From 1973 through 1975, each Service estimated an AFQT score from its own test battery. The ASVAB became operational as the single DoD enlistment test in 1976, and AFQT scores have been based since then on a common test. The AFQT composite of ASVAB used in this study (Form 8A) was calibrated against an earlier version of AFQT (Form 7A) used operationally from 1960 through 1972. This calibration established the linkage to the World War II reference population, thereby enabling percentile scores from the new AFQT to have the same interpretive meaning as scores from predecessor tests.

AFQT Categories. For reporting purposes, AFQT scores have traditionally been grouped into five broad categories. Persons whose scores place them in Categories I and II are above average in trainability; those in Category III, average; those in Category IV, below average; and those in Category V, markedly below average and, under current Service policy, not eligible to enlist. The Services prefer enlistees in the higher AFQT categories because training time and associated costs are

lower, and such recruits are more likely to qualify for specialized training in a greater number of occupational areas. Table 2 shows the percentile scores for the various categories and the percent of the World War II reference population in each. AFQT percentile scores are based on the World War II population of officers and enlisted men who were on active duty as of December 31, 1944 — 11,694,229 males.

Table 2

Armed Forces Qualification Test (AFQT) Categories by Corresponding Percentile Score Range and Distribution of World War II Reference Population

AFQT Category	Percentile Score Range	World War II Reference Population Distribution (Percent)
I	93 - 100	8
II	65 - 92	28
III	31 - 64	34
IV	10 - 30	21
V	1 - 9	9
		100

STUDY METHODOLOGY

The Profile of American Youth is closely related to the five-year National Longitudinal Survey of Youth Labor Force Behavior (NLS) in three ways. The first and most important relationship is that the profile study used for its sample young people who completed the first annual interview of the NLS in 1979. The profile study used the NLS sample because it was an already-existing nationally representative sample of young people in the age group of interest. Second, the data collection for both studies was carried out by the National Opinion Research Center (NORC). Third, there would be a sharing of data between the two studies. Demographic data collected by the NLS were added to the ASVAB test information obtained in the profile study.

The purpose of the NLS is to study the behavior within the labor market of a large and representative cross-section of American youth. Information about youth born from 1957 through 1964 is being collected through annual personal interviews. The NLS is primarily concerned with problems relating to employment and unemployment. The interviews also gather a great deal of supplemental information about the characteristics, experience, plans, and attitudes of the young people.

STUDY RESEARCH DESIGN

The Sample. The NLS sample was designed to represent the national population of youth ages 14 to 21 as of January 1, 1979 (Frankel & McWilliams, 1981; McWilliams & Frankel, 1981). Civilian members of the youth population were obtained by screening approximately 80,000 households during the fall of 1978. This screening identified approximately 14,000 eligible youth of the appropriate age. Members of the youth population serving in the military were selected in the fall of 1978 from lists provided by the Defense Manpower Data Center (DMDC). Youth in the military were eligible for selection if they were (a) serving in the Armed Services as of September 30, 1978 and (b) would be between the ages of 17 and 21 as of January 1, 1979. In the spring of 1979, NORC interviewed 12,686 civilian and military youth for the first annual (baseyear) NLS survey.

The NLS baseyear sample contains youth from both urban and rural areas, youth from all major census divisions, and approximately equal proportions of males and females. The sample overrepresents, in a statistically appropriate way, certain key groups, such as Hispanics, blacks, economically disadvantaged whites, and women in the military. This overrepresentation allows for more precise analyses of these groups than would otherwise be possible.

The profile study used for its target sample the 12,686 young people who completed the first annual (1979) interview of the NLS. The 11,914 tests administered represent a completion rate of approximately 94 percent. Table 3 shows the composition of the completed profile sample by sex and race/ethnicity.

Table 3

Composition of the Profile of American Youth Sample: Racial/Ethnic Group and Sex

Racial/Ethnic Group	Sex		Total
	Male	Female	
White[a]	3,531	3,496	7,027
Black[b]	1,511	1,511	3,022
Hispanic	902	927	1,829
Total	5,944	5,934	11,878

[a]White includes all racial/ethnic groups other than black or Hispanic.
[b]Black does not include persons of Hispanic origin.

Since the Services primarily recruit individuals who are 18 years of age an older, the profile study focused upon young people born between January 1, 1957 and December 31, 1962. Thus, the age range for the profile study sample is 18 through 23 years at the time of testing. Table 4 shows the profile study sample of 9,173 people of enlistment age. Table 5 displays the corresponding size of the 1980 national youth population (weighted sample) by year of birth, race/ethnicity, and sex.

Table 4

Composition of the Profile of American Youth Sample: Year of Birth, Racial/Ethnic Group, and Sex[a]

Year of Birth	Age at Time of Testing (Years)	Racial/Ethnic Group								
		White[b]		Black[c]		Hispanic		Total		
		Male	Female	Male	Female	Male	Female	Male	Female	Total
1962	18	458	481	213	210	108	145	779	756	1,535
1961	19	383	418	207	211	129	116	989	745	1,444
1960	20	445	448	197	206	123	110	765	764	1,529
1959	21	490	519	185	195	108	109	767	823	1,590
1958	22	477	505	190	167	92	102	759	774	1,533
1957	23	521	488	167	166	93	107	781	761	1,542
TOTAL		2,754	2,779	1,143	1,155	653	689	4,550	4,623	9,173

[a]Restricted to persons in the sample born between January 1, 1957 and December 31, 1962 (18 through 23 years at time of testing, July-October 1980).
[b]White includes all racial/ethnic groups other than black or Hispanic.
[c]Black does not include persons of Hispanic origin.

Table 5

Composition of National Youth Population Based on Profile of American Youth Sample: Year of Birth, Racial/Ethnic Group, and Sex[a]

(In Thousands)[b]

Year of Birth	Age at Time of Testing (Years)	Racial/Ethnic Group								
		White[c]		Black[d]		Hispanic		Total		
		Male	Female	Male	Female	Male	Female	Male	Female	Total
1962	18	1,677.9	1,616.1	296.4	292.1	138.5	123.5	2,112.8	2,031.7	4,144.5
1961	19	1,701.8	1,643.9	296.8	293.1	140.0	124.3	2,138.2	2,061.3	4,199.5
1960	20	1,729.8	1,688.8	295.9	290.2	134.8	127.8	2,160.1	2,087.8	4,248.0
1959	21	1,753.2	1,675.3	285.2	289.3	120.1	131.8	2,158.8	2,096.4	4,255.1
1958	22	1,758.5	1,708.7	284.1	289.5	122.0	131.7	2,161.6	2,129.9	4,291.4
1957	23	1,762.8	1,708.4	275.7	282.9	121.2	127.5	2,159.7	2,110.8	4,270.4
TOTAL		10,388.6	10,014.2	1,733.0	1,737.1	777.6	766.6	12,891.2	12,517.9	25,408.1

[a]Restricted to persons in the sample born between January 1, 1957 and December 31, 1962 (18 through 23 years at time of testing, July-October 1980).
[b]Figures are rounded.
[c]White includes all racial/ethnic groups other than black or Hispanic.
[d]Black does not include persons of Hispanic origin.

Quality of the Sample. To provide DoD with an assessment of the sample design, development of sample case weights and sampling statistics, an independent panel of sampling experts (Dr. B. F. King, University of Washington; Dr. L. Kish, University of Michigan; Dr. G. E. Hall, U. S. Bureau of Census; and Dr. J. Sedransk, State University of New York) was convened. The panel concluded: (a) the sample design was appropriate for meeting the objectives of the profile study and (b) all of the statistical procedures used in the development of sample case weights and sampling statistics met the professional criteria established for efforts of this nature, both in the public and private sectors. (Frankel & McWilliams 1981).

TEST ADMINISTRATION

During the period July through October 1980, NORC representatives administered the ASVAB to the 11,914 young people who comprise the profile sample. Testing was generally conducted in groups of five to ten persons. More than 400 test sites, including hotels, community centers, and libraries throughout the United States and abroad were used. The test was administered according to strict guidelines conforming to ASVAB procedures, which assured both accuracy and consistency of results. Great care was also taken to assure confidentiality.

In May 1981, NORC sent to all respondents a copy of their test results, information to interpret the scores, and a brochure containing vocational and educational information. In addition, participants were paid honoraria for completing the test. The decision to pay an honorarium was based on experience in similar studies which indicated that a powerful incentive would be needed in order to get young people to travel up to an hour to a testing center, spend three hours or more taking a test, and then travel home. The honorarium was set at $50.00.

NORC's decision to provide an incentive honorarium was also influenced by the importance of the NLS and an obligation to ensure that the added demands of the profile study on the NLS respondents would do nothing to damage further NLS participation. It was anticipated that the monetary incentive offered for participation in the aptitude profile study would work against attrition of the NLS sample and would even increase the goodwill of its members.

STUDY QUALITY CONTROL

Quality of Data Files. A DoD team of testing experts and computer programmers verified that ASVAB scores and demographic information had been accurately transcribed from the original source documents (i.e., answer sheets and questionnaires) to the computer tape provided to DoD. A random sample (one percent of the cases) was selected for the data audit. For the sample cases, ASVAB answer sheets were hand-scored and demographic questionnaires were manually reviewed. In every case, the information from the source documents had been correctly recorded (Sellman & Hagan, 1981).

Quality of ASVAB. To evaluate the suitability of the ASVAB for measuring the aptitudes of a national sample of young people, DoD contracted with Dr. R. D. Bock, an authority on educational and psychological testing from the University of Chicago. Dr. Bock evaluated the test to determine its appropriateness for measuring vocational aptitudes and its equity for minorities and females. He concluded that the ASVAB is useful for measuring aptitudes of civilian youth and that cultural test bias was not apparent for minorities and females. Moreover, he indicated that the quality of ASVAB equals or surpasses that of commercial aptitude and achievement tests (Bock & Mislevy, 1981).

REFERENCES

Bock, R. D. and Mislevy, R. J. Data Quality Analysis of the Armed Services Vocational Aptitude Battery. Chicago, IL: National Opinion Research Center, August 1981.

Frankel, M. R. and McWilliams, H. A. The Profile of American Youth: Technical Sampling Report. Chicago, IL: National Opinion Research Center, March 1981.

Ginzberg, E., Anderson, J. K., Ginsburg, S. W., and Herma, J. L. The Lost Divisions. New York: Columbia University Press, 1959.

McWilliams, H. A. The Profile of American Youth: Field Report. Chicago, IL: National Opinion Research Center, December 1980.

McWilliams, H. A. and Frankel, M. R. The Profile of American Youth: Non-Technical Sampling Report. Chicago, IL: National Opinion Research Center, October 1981.

Sellman, W. S. and Hagan, H. T. The Profile of American Youth: Data Audit. Technical Memorandum 81-1. Washington, D. C. Directorate for Accession Policy, Office of the Secretary of Defense, April 1981.

Sheatsley, P. B. The Profile of American Youth: Pretest Report. Chicago, IL: National Opinion Research Center, September 1980.

Uhlaner, J. E. and Bolanovich, D. J., Development of the Armed Forces Qualification Test and Predecessory Army Screening Tests, 1946-1950. PRS Report 976. Washington, D. C.: Personnel Research Section, Department of the Army, November 7, 1952.

MILITARY AND CIVILIAN TEST SCORE TRENDS (1950-1980)

Brian K. Waters
Mark J. Eitelberg
Janice H. Laurence

Since 1975, the College Entrance Examination Board (CEEB) has published several reports on Scholastic Aptitude Test (SAT) score decline. The CEEB data were similar to data on the American College Testing (ACT) Program, as well as a number of achievement tests. The subject of declining student aptitudes and achievement has since dominated space in educational and psychological literature, with many reports and books receiving heavy media and public exposure. Numerous symposia, commissions, and studies have also been launched to answer three key questions: 1) Are the test score declines a "real" national phenomenon?; 2) What are the cause(s) for the decline?; and 3) What can be done to reverse the trend of declining scores?

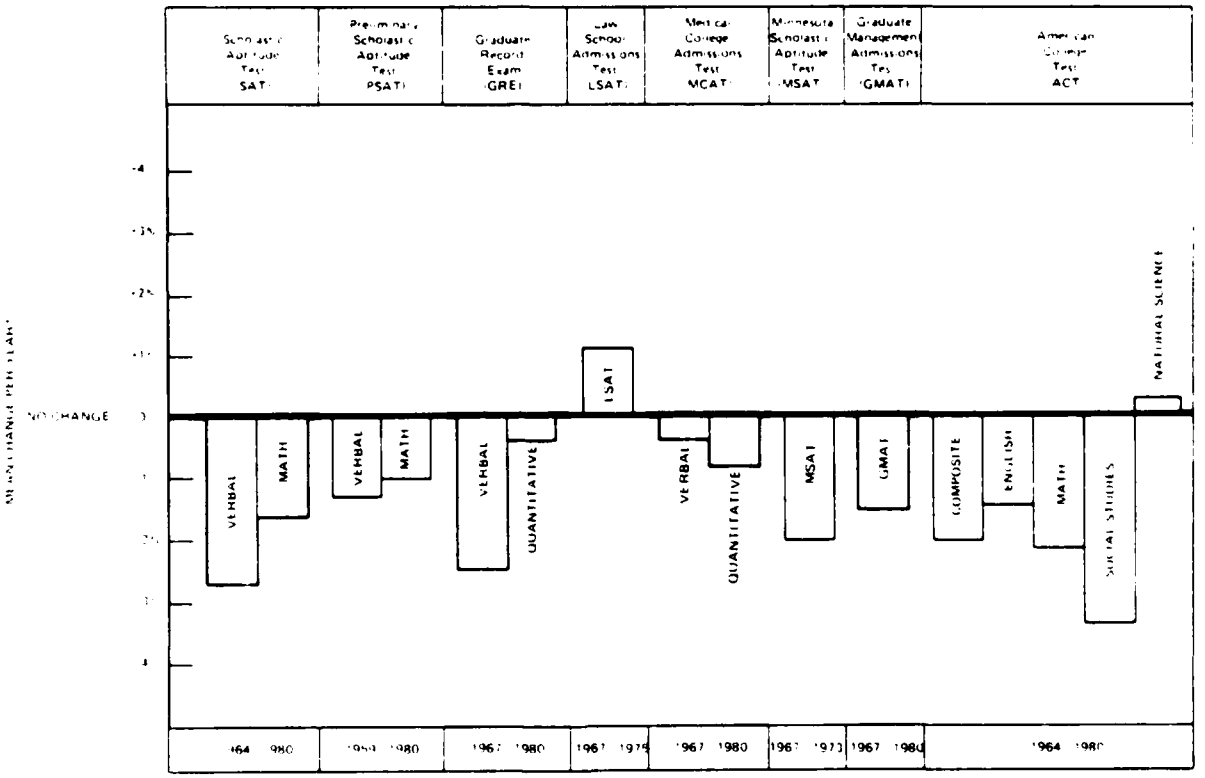
The nature and scope of aptitude and achievement test score changes in the national population are of considerable interest to military manpower and personnel managers. The national population provides the pool from which military applicants are drawn (also draftees during periods of national emergency). The civilian population also provides a baseline upon which to assess current and historical recruit quality. And, in the context of the current major research effort to profile the aptitudes of American youth, a review of the civilian aptitude and achievement test score decline places into better perspective military test score trends over the survey period.

This paper describes test score changes between the early 1950s and 1980 (when the profile study was conducted). The paper begins with a tabular and graphic picture of national scholastic aptitude and achievement test score trends. The paper then describes military accession test score trends on AFOT from 1967 to 1980. AFQT score trends are reported by high, median, and low scores. The paper concludes with a brief summary and interpretation of the trends.

CIVILIAN SCHOLASTIC APTITUDE TEST SCORE TRENDS

The most clear, consistent, and unambiguous evidence of the decline in the civilian target population comes from the aptitude testing domain. Table 1 and Figure 1 provide a compilation of the trends in this country since the early 1950s.

The aptitude test data show remarkable consistency. With the exception of slight increases on the Medical College Admissions Test-Quantitative Subtest and the Law School Admissions Test, the other measures of scholastic aptitude consistently decreased at a rate of about one to three percent of a standard deviation per year. This trend continues, although there is some evidence that the rate of decline has slowed somewhat through 1980. Other major trends show that verbal scores have tended to decrease faster than quantitative scores; female scores have declined more rapidly than male scores, particularly in the verbal domain; and overall aptitude test scores increased from 1950 through about 1965, and decreased consistently until the late 1970s. There appears to be a lessening of the rate of decline between 1977 and 1980, although the decline continued through 1980.[1] The "causes" of the consistent aptitude test score patterns are not at all clear. Nevertheless, the general conclusion of most authors is that there are multiple factors contributing to the trend.



Source: (Waters, 1981)

Figure 1. Civilian Aptitude Measures

[1] A recent release from ETS reports that 1981 mean SAT scores were identical to 1980 mean scores on both quantitative and verbal subtests.

Table 1

Scholastic Aptitude Measures

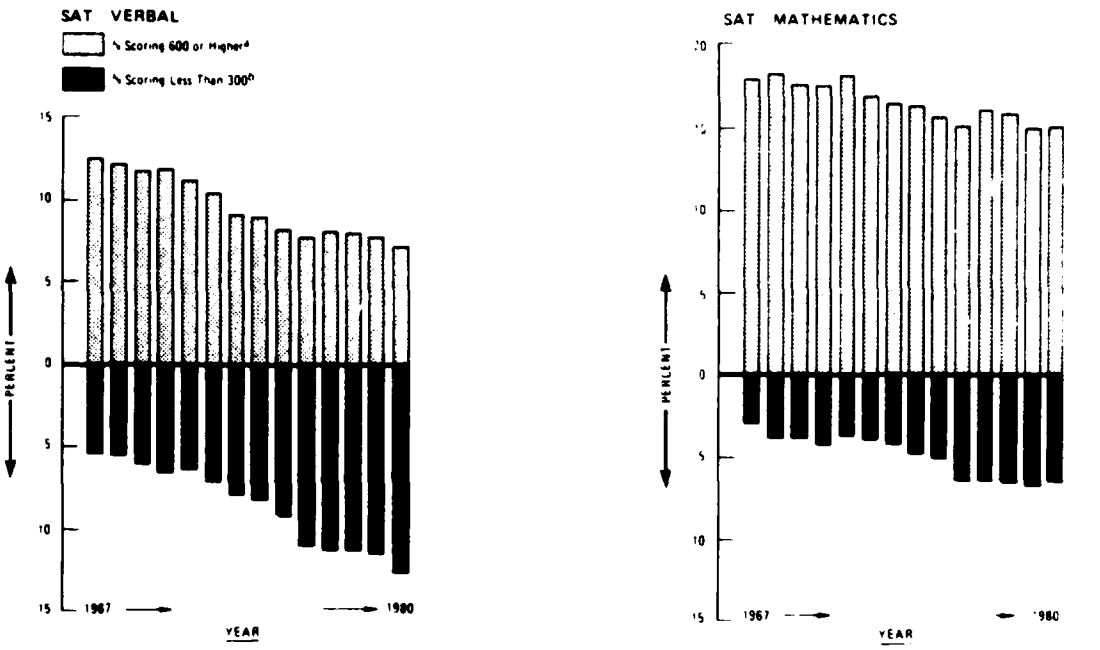
						Trends (SD YR)
Scholastic Aptitude Test (SAT)	Verbal	1967 1980	11 12	1 500 000	NE E EC	Male 2 3
		1967 1980				Female 3 2
		1952 1963				Total -0 2
		1964 1980				Total 2 7
	Mathematics	1967 1980	11 12	1 500 000	NE E EC	Male 1 4
		1967 1980				Female 1 6
		1952 1963				Total -0 6
		1964 1980				Total 1 6
American College Test (ACT)	Composite	1964 1980	11 12	850 000	NC S W	Male 1 3
						Female 2 5
						Total -2 0
	English	1964 1980	11 12	850 000	NC S W	Male -0 9
						Female -2 3
						Total -1 4
	Mathematics	1964 1980	11 12	850 000	NC S W	Male 2 3
						Female -2 0
						Total -2 3
	Social Studies	1964 1980	11 12	850 000	NC S W	Male -2 4
						Female -4 1
						Total 3 3
	Natural Science	1964 1980	11 12	850 000	NC S W	Male -0 9
						Female -0
						Total -0 3
Preliminary Scholastic Aptitude Test (PSAT)	Verbal	1958 1980	11	1 000 000	NE E EC	Male 0 7
						Female 1 8
						Total -1 3
	Mathematics	1958 1980	11	1 000 000	NE E EC	Male 1 1
						Female -0 8
						Total 1 0
Minnesota Scholastic Aptitude Test[2] (MSAT)	Form A	1958 1966	11	60 000	Minn	Total -6 2
	Form C	1967 1973	11	65 000	Minn	Total -2 0
Graduate Record Exam (GRE)	Verbal	1967 1980	16	800 000	National	Total 1 3
	Quantitative	1967 1980	16	Not Available	National	Total 0 4
Law School Admissions Test (LSAT)		1967 1975	16	Not Available	National	Total -0 6
Medical College Admissions Test (MCAT)	Verbal	1967 1975	16	55 000	National	Total -1 8
		1977 1980				Total -2 6
	Quantitative	1967 1975	16	55 000	National	Total +1 0
		1977 1980				Total -4 3
Graduate Management Admissions Test (GMAT)		1967 1975	16	400 000	National	Total -1 5

[1] PSAT Scores corrected for scale drift 1967 1980

[2] MSAT Calculations split in 1966 1967 when current form was introduced No MSAT after 1973

Source: (Waters, 1981).

SAT Score Trends. Figure 2 displays SAT Verbal and Mathematics subtest trends from 1967 to 1980.



[a]Scores of 600 on the SAT VERBAL and the SAT MATH equate approximately to scores in the 90th percentile and 92nd percentile respectively on the AFQT distribution

[b]Scores of 300 on the SAT VERBAL and the SAT MATH equate approximately to scores in the 50th percentile and 31st percentile respectively on the AFQT distribution

SOURCES College Entrance Examination Board

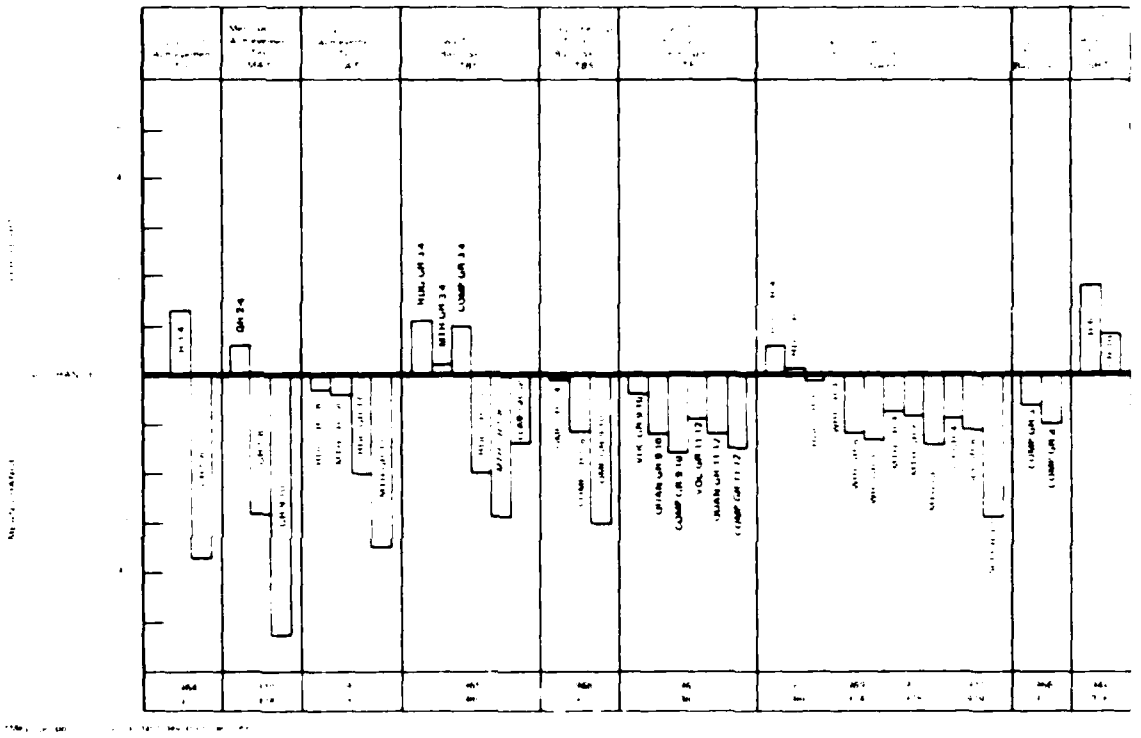
Figure 2. **Score Distributions on the Scholastic Aptitude Test (SAT), 1967–1980**

Figure 2 depicts the relatively consistent decline on both subtests at both ends of the scale range. As will be seen in the next section of this paper, these declines are similar to test score patterns of military recruits who score at the higher and lower ends of the Armed Forces Qualification Test (AFQT).

CIVILIAN SCHOLASTIC ACHIEVEMENT TEST SCORE TRENDS

Table 2 and Figure 3 depict 1964-1980 mean results for 10 achievement test batteries. Data for the individual batteries have been grouped, when available, into grades 1-4, 5-8, and 9-12 and by subtests that roughly parallel the verbal/quantitative/composite breakouts of the aptitude measures discussed above. As might be expected, long-term trends across achievement content areas are not as consistent as across the more "factorially pure" aptitude areas. Trend data are displayed in Table 2 by percent change in standard deviation per year where both means and variances were provided in the original source. Figure 3 displays Table 2 graphically.

In general, the authors found consistent evidence of achievement test score declines in all areas tested above grade 4, for the 1960s through the 1970s. Pre-schoolers and children in the early years of grammar school (1st - 3rd grade) generally scored higher on all measures, while the scores of 4th grade students remained fairly stable. In the opinion of the authors, these trends are real, national in scope, and continuing -- though at a decreasing rate of decline since about 1977.



Source: (Waters, 1981)

Figure 3. Civilian Achievement Measures

OTHER INDICATORS OF POPULATION PERFORMANCE CHANGE

The literature on the test score decline includes many references to other indicators of a declining national level of academic competence of *youth. These indicators include elementary, high school, and college* teachers' opinions, statewide competency-based assessment, measures of curricula content at all levels, analyses of classroom hours attendance per student, analyses of teacher education and practices, and physiological hypotheses about diet, drug, medication, nuclear radiation and other possible correlates of declining test scores. It is beyond the scope of this paper to attempt to analyze the probable or possible causes of the declining scores; however, an excellent review by Rimland and Larson, (1980) is *available.*

Table 2

Scholastic Achievement Measures

Instrument		Time Period	Grade(s)	Annual N	Area	Trends % SD / Yr
Stanford Achievement Test		1964 1973	1 4	6 000 000	National	-1 3
			5 8			3 7
Metropolitan Achievement Test (MAT)		1970 1978	2 4	400 000	National	-0 6
			5 8			2 8
			9 10			-5 3
California Achievement Test (CAT)		1970 1978	2	Not Available	National	Slight Gain
			5 8			Rdg 0 3
						Mth 0 4
			11			Rdg 2 0
						Mth 3 5
Iowa Tests of Basic Skills (ITBS)		1965 1980	3 4	50 000	Iowa	Rdg -1 1
						Mth -0 2
						Comp -1 0
			5 8			Rdg 2 0
						Mth -2 9
						Comp 1 4
Comprehensive Tests of Basic Skills (CTBS)		1968 1973	2 4	200 000	National	Comp 0 1
			5 8			Comp 1 2
			9 10			Comp 3 0
Iowa Tests of Educational Development (ITED)		1965 1980	9 10	Not Available	Iowa	Voc 0 4
						Quan 1 2
						Comp 1 6
			11 12			Voc 0 9
						Quan 1 2
						Comp -1 5
National Assessment of Educational Progress (NAEP)	Rdg	1971 1980	4	75 000 to 100 000	National	Rdg + 06
						Wrt 0
	Mth	1969 1974				Mth 0 7
						Sci 0 9
	Mth	1973 1978	8			Rdg -0 1
						Wrt 1 2
	Sci	1970 1979				Mth 0 8
						Sci -1 1
			12			Rdg 0 1
						Wrt -1 3
						Mth 1 4
						Sci -2 9
Canadian Tests of Basic Skills		1966 1973	3	Not Available	Canada	Comp -0 6
			8			Comp 1 8
Iowa Silent Reading Tests (ISRT)		1944/1976	6	15 000/0 000	Indiana	-1 8
			10	11 000/0 000		-0 8
General Educational Development (GED)		1964 1979	Mean	120 000 to 700 000	National	-0 05 % Net Stds/Yr 173% Vs 69 1%
			16			

[1] ISRT scores adjusted for examinee age changes between scoring periods

Source: (Waters, 1981)

SUMMARY OF OVERALL CIVILIAN TESTING TRENDS

It is evident that national youth performance on scholastic aptitude and achievement tests has been in a state of decline. Assuming comparability of populations (for current military-age youth, between the ages of 17-24), the scope of the decline would likely represent a decrease of about one-fifth to one-third of a standard deviation on the average from the 1970 pool of AFQT examinees, or about 2-3 percent of a standard deviation per year. This rate would equate to a decline of approximately 4-5 raw score points for the average military enlisted recruit between FY 1971 and FY 1980. The next section of this paper looks at the military data.

MILITARY RECRUIT APTITUDE TEST SCORE TRENDS

Figure 5 displays the percentages of non-prior service military accessions who scored in AFQT Category I (Sellman & Laurence, 1981) between FY 1967 through FY 1980. Category I scores are roughly equivalent to a score of 600 and above on the SAT (shown in Figure 2 above).

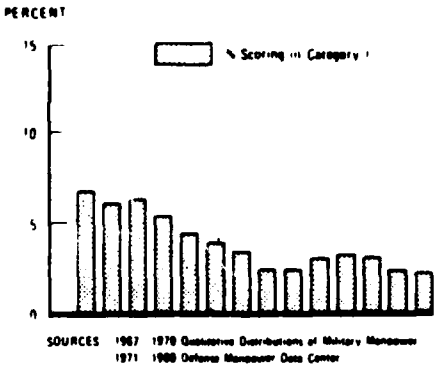


Figure 4. Military Accessions: AFQT Category I (Top 8%) 1967–1980

AFQT Category IV military accession statistics are strongly influenced by DoD and Service policies, changes in the recruiting and job market, and other factors independent of the aptitude levels in the national population of enlistment-age youth. Nevetheless, the proportion of AFQT Category IV accessions did increase from 23 percent in FY 1967 - FY 1969 to 26 percent in FY 1979 - FY 1981. The latter figure probably reflects the effects of the miscalibration of AFQT for the first two years of the latter period, (Department of Defense, July 1980). However, FY 1981 test scores are correct. The three percent increase from the earlier period is roughly comparable to the increased percentages in SAT Verbal and Mathematics scores below 300 over the same time frame (Figure 2).

A median provides a single index of the full distribution of AFQT scores. The medians declined from 79.4 to 72.4 or seven AFQT raw score points over the 14-year period. This decline parallels the two to three percent standard deviation yearly decline observed for civilian aptitude test scores during the same year period. Figure 5 shows median AFQT scores grouped in three-year periods from FY 1967 through FY 1981.

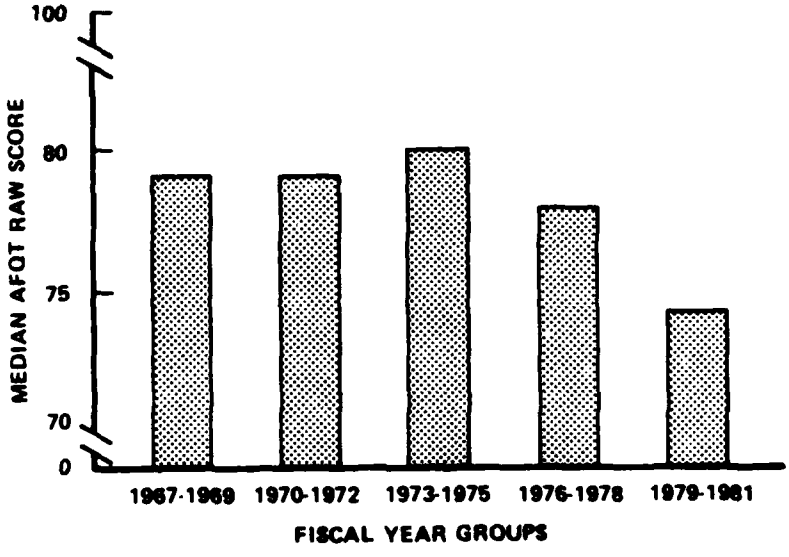


Figure 5. Median AFQT Raw Scores for Military Non-Prior Service Accessions, 1967-1981, by Three Year Periods

IMPLICATIONS OF CIVILIAN AND MILITARY RECRUIT APTITUDE TEST SCORE TRENDS

It appears that civilian aptitude test score trends provide useful information to Defense manpower planners on the aptitude levels of American youth in the pool from which potential recruits are drawn. It is therefore important that the Department of Defense monitor national test score trends for both short-term recruiting and mobilization planning purposes.

REFERENCES

Department of Defense. Aptitude Testing of Recruits. A Report to the House Committee on Armed Services. Washington, D.C.: Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs and Logistics), July 1980.

Rimland, B. and Larson, G. E. The Manpower Quality Decline: An Ecological Perspective. NPRDC Technical Note 81-84. San Diego: Navy Personnel Research and Development Center, November 1980.

Sellman, W. S. and Laurence, J. H. "Aptitude Testing in DoD and the Profile of American Youth Study." A paper presented at the 23rd Annual Conference of the Military Testing Association, Washington, D. C., October 1981.

Waters, B. K. The Test Score Decline: A Review and Annotated Bibliography. Technical Memorandum 81-2. Washington, D.C.: Directorate for Accession Policy, Office of the Secretary of Defense, August 1981.

SUBPOPULATION ANALYSES OF CURRENT YOUTH APTITUDES [1]

by

Mark J. Eitelberg
Janice H. Laurence
Brian K. Waters

Human Resources Research Organization

and

Wayne S. Sellman
Office of the Assistant Secretary of Defense
(Manpower, Reserve Affairs and Logistics)

In 1980, the National Opinion Research Center (NORC) of the University of Chicago administered the Armed Services Vocational Aptitude Battery (ASVAB) to a national probability sample of approximately 12,000 young men and women. The procedure and methods used to select the sample were designed to yield a data base of youth that could be statistically projected (within known confidence intervals) to represent the entire population (and important subgroups) born in 1957 through 1964 (Frankel & McWilliams, 1981).

The results will be analyzed to identify subpopulation differences in test performance and to determine qualification rates for military service. The demographic variables selected for analysis are age, sex, race/ethnicity, level of education, socioeconomic status, and geographic region. Subpopulation comparisons will be made on the basis of the Armed Forces Qualification Test (AFQT), four aptitude composites, and an estimate of reading ability. The subpopulation analyses have not been completed. Therefore, this paper contains only a description of the background, methodology, and scope of the demographic comparisons.

COMPARISON MEASURES

Mean AFQT percentile scores will be used since AFQT results are typically reported in terms of this metric. The raw AFQT scores of individuals will be converted to AFQT percentile scores, and the mean percentile scores for each subgroup will then be calculated.

The four ASVAB aptitude composites selected for analysis are Mechanical (M), Administrative (A), General (G), and Electronics (E). The ASVAB subtests comprising the Administrative, General, and Electronics composites are the same in all four Services. The Air Force version of the Mechanical composite was used for the subpopulation analyses. The individual subtests that comprise these composites are shown in Table 1.

---

[1] A paper presented at the 23rd Annual Conference of the Military Testing Association, Washington, D.C., October 27, 1981. The views expressed in this paper represent those of the authors and do not necessarily reflect the policy or opinion of the Department of Defense.

Common Aptitude Composites and Their Component ASVAB Subtests

(Forms 8, 9, and 10)

Common Aptitude Composites	ASVAB Subtests
Mechanical (M)[a]	Mechanical Comprehension Automotive-Shop Information General Science
Administrative (A)	Coding Speed Numerical Operations Paragraph Comprehension Word Knowledge
General (G)	Arithmetic Reasoning Paragraph Comprehension Word Knowledge
Electronics (E)	Arithmetic Reasoning Electronics Information General Science Mathematics Knowledge

[a]The Administrative, General, and Electronics composites are the same for all four Services. For the purpose of population subgroup analyses, this report uses the Air Force version of the Mechanical composite.

Table 1

Estimates of reading ability will be obtained for the profile study subgroups by converting ASVAB General composite scores to comparable scores on the Adult Basic Learning Examination (ABLE) (see Mathews, Valentine, & Sellman, 1978). ABLE is a battery of tests (vocabulary, spelling, reading, arithmetic/computation, and arithmetic/problem solving) designed to measure the educational achievement of adults who have not completed high school. ABLE covers 12 years of school achievement through the use of three separate levels of test batteries. Since the ASVAB General composite (which combines Paragraph Comprehension, Word Knowledge, and Arithmetic Reasoning subtests) correlates so highly (.85) with ABLE, it will be possible to convert the General composite scores to ABLE scores, and then use these measures as estimates of reading ability expressed in terms of scholastic grade levels.

SUBPOPULATION ANALYSES

AGE

Background. The Army Alpha tests from World War I provided some of the first documented evidence of population differences based on chronological age. Since that time, numerous cross-sectional studies have supported the finding that mental ability (a) reaches a peak in early adulthood (the mid-twenties); (b) declines gradually to about age 50; and (c) drops sharply thereafter. Longitudinal studies conducted since the early 1950s, however, indicate that the pattern of intellectual growth and decline is somewhat different from that which is found in cross-sectional research. Although there is still little longitudinal evidence concerning the shape of the so-called "age curve," the data now imply (a) a pattern of intellectual growth through early adulthood; (b) general stability during the middle decades of life (with increases in certain abilities and decreases in others); (c) a gradual and minor decline beginning after the age of 50; and (d) increased decline during the 70s and 80s.

Two-year groupings will be used to separate the 1980 youth population by age. The age categories, by year of birth and age at time of testing, are as follows: 1961 and 1962 (ages 18 and 19); 1959 and 1960 (ages 20 and 21); 1957 and 1958 (ages 22 and 23).

The analysis of age differences will concentrate on mean AFQT percentile scores and measures of reading ability.

SEX

Background. Many standardized tests of general aptitude are designed to eliminate (or counterbalance) items or subtests that result in systematically higher scores for one sex over the other. The belief that differential factors should be minimized or balanced is based on the assumption that (a) there is no clear understanding of which specific test items are the best indicators of general aptitude and (b) no special "advantage" in measured performance on these tests should be given to either sex.

Nevertheless, the consistent trend has been that males tend to excel on tests of mathematical reasoning (or quantitative ability), spatial abilities, and mechanical/science aptitudes; and females tend to excel on tests involving verbal fluency or the mechanics of language, memory abilities, perceptual speed, and manual dexterity (Tyler, 1965; Maccoby & Jacklin, 1974).

The AFQT measures verbal and quantitative abilities in approximately equal proportion. This balance reduces the likelihood of sex-related differences in test performance.

The analysis will present data on the mean AFQT percentile scores of males and females by the three age groups. Mean percentile scores of males and females on the four aptitude composites and the estimated reading grade levels of young men and women in the general population will also be presented.

RACE/ETHNICITY

Background. In this country, most studies of racial/ethnic group test performance focus primarily on the differential abilities of white and black children and young adults. Published evidence suggests that, on standardized tests of mental ability, (a) whites, on the average, score higher than blacks; (b) average group differences remain fairly constant during the school years (the smallest differences occur at the very young ages); (c) blacks perform relatively better on verbal tests than on non-verbal tests; (d) the socioeconomic, geographic, and educational correlates for racial minority groups and whites are generally similar (though there are some differences in the magnitude of correlation); and, further, (e) the differences between individuals of the same race exceed in magnitude the average differences between separate races.

Attempts to measure racial differences in test performance in the civilian sector can be traced back as far as the late nineteenth century. As in the military testing experience (Eitelberg, 1981), there is a remarkable unanimity of results in civilian testing: at each age-level and under a variety of conditions, blacks, on the average, regularly score below whites (Jensen, 1980; Scarr, 1981). There are regional variations; nevertheless, these variations are similar for blacks and whites, and the racial differential remains fairly constant from one region to another.

Although the majority of studies involving racial/ethnic groups in this country currently concentrate on the differences between whites and blacks, there is a long history of research regarding the relative abilities of different "ethnic" (i.e., national origin) groups. The volume of scientific research on the topic of ethnic differences (or race differences other than those between whites and blacks) has lessened greatly since World War II. Nevertheless, there is some degree of consistency in the data on test performance. For instance, study results in this country show that (a) white school children of European ancestry score, on the average, considerably higher than children from racial and ethnic minority groups (with the notable exception of certain Asian-American groups) — and especially those that are socioeconomically disadvantaged (e.g., Hispanics, Native Americans, as well as blacks); and (b) the test performance of racial/ethnic groups oftens varies with respect to the particular type of test used (that is, some groups perform noticeably better on certain kinds of tests than on others, and the extent of group differences will change according to the types of tests that are emphasized).

The profile study results classify the population into three groups: white and others (including all non-Hispanic and non-black racial/ethnic subgroups), black (non-Hispanic), and Hispanic. These three groups are used since they represent the largest relative racial/ethnic subgroups within the general population. Yet, it should be noted that the Hispanic category includes several separate ethnic groups (e.g., Mexican-Americans, Puerto Ricans, Cubans and other Latin Americans, Spanish and Portuguese) variously described simply as being of "Hispanic" origin. Furthermore, the category defined as "white and others" includes Native Americans, Pacific Islanders, and persons of Asian ancestry. (Since the data are weighted, and the proportion of "non-white" groups in the general population is so small in comparison with whites, the differences between the combined group and a "white only" group are negligible.) For the purposes of the profile study, then, references to the "white" and "white and other" racial/ethnic groups are synonymous.

The mean AFQT percentile scores for whites, blacks, and Hispanics will be analyzed (by total subgroup and two-year age categories). In addition to the AFQT score comparisons, the mean score for males and for females within the separate racial/ethnic subgroups will be compared on the basis of common aptitude composites and estimated reading ability.

LEVEL OF EDUCATION

Background. There is a strong positive correlation between aptitude test performance and the amount of formal education. There are, however, several problems involved in using years of schooling as a focus of analysis. For example, there are differences in the quality of education from geographical region to region, school to school, and other related factors. In addition, education variables are not easily isolated or separated from other variables (e.g., age and socio-economic status).

For the profile study, educational attainment is defined according to high school graduation status. The three categories of graduation status are: (a) non-high school graduate (including, in some cases, high school students as well as drop-outs); (b) recipient of the General Educational Development (GED) high school equivalency certificate; and (c)

high school diploma graduate (also including all persons, regardless of high school graduation status, with education at the college level).

Both AFQT and aptitude composite percentile scores for the three education categories will be analyzed. Composite score comparisons will include the mean scores for high school graduates, GEDs, and non-high school graduates by sex.

SOCIOECONOMIC STATUS

Background. Social class differences have been reported in numerous studies from the earliest days of psychological testing. During World War I, average scores on the Army Alpha test demonstrated a clear relationship to preservice employment. Highest scores were obtained by those in professional occupations (e.g., engineer, accountant), ranging down to those who had worked as unskilled laborers (in preservice jobs) at the very bottom of the scale. Studies of Army General Classification Test (AGCT) scores from World War II revealed a similar pattern of test performance by soldiers according to preservice occupational categories (Anastasi, 1958; Tyler, 1965).

When children are classified on the basis of their father's occupation, the same sort of differentiation in test scores is apparent. Children of parents in the professions generally score highest on aptitude tests, and children of day-laborers and unskilled workers generally score lowest (Anastasi, 1958).

In general, studies that have examined social class differences are consistent. Adults and children from more-privileged homes perform better, on the average, than do those from less-privileged homes. The relationship between socioeconomic status (SES) and performance on tests of mental ability is thus one of the most consistent and least questioned outcomes of standardized testing (Tyler, 1965).

The SES of children and adolescents is typically indexed using mother's education, father's education, average family income, and father's occupational status. None of these four variables alone explains all of the variation in ability attributable to "family background." Nevertheless, there is a strong correlation between the variables, and research has shown that each affects intellectual ability in a different manner but to a similar degree (Sewell and Hauser, 1975; Featherman, 1980). Recent analyses suggest that the measured effects of mother's education on ASVAB performance approximate the measured effects of all four variables combined (Bock and Moore, 1981). For the 1980 youth population, then, mother's education is used in place of an SES index as a general indicator of family background.

The mean AFQT percentile scores of the 1980 youth population, arranged according to five categories of mother's education, will be considered. The five categories are as follows: 8th grade or less; grades 9-11; high school graduate; some college; college graduate or above.

GEOGRAPHIC REGION

Background. Regional variations in test performance have been commonly reported. Generally, average scores on tests of mental ability are lowest in the deep South; average test socres increase in almost gradient fashion to the North and West.

Of course, regional differences are greatly affected by factors related to urban and rural environments. The preponderance of evidence on urban-rural differences shows that persons from rural regions receive lower average scores on tests of mental ability than do persons from urban regions. A combination of causes may be at work (including differences in the quality of education, socioeconomic variations, and cultural factors), but the same results have been found repeatedly in a wide variety of studies in many parts of this country as well as in Europe.

The mean AFQT percentile scores of individuals residing in different geographic regions (the nine "divisions" used by the Bureau of Census) will be compared and evaluated.

ELIGIBILITY FOR ENLISTMENT

Each Military Service applies its own aptitude standards in determining eligibility for enlistment. As seen in Table 2, these aptitude standards vary according to educational attainment (high school graduation status) and, in the Navy and Marine Corps, according to sex. For example, in the Army, male and female high school graduates are currently required to achieve a minimum AFQT percentile score of 16 and a standard score of at least 85 on one of nine Service-specific aptitude composites. In contrast, Air Force enlistment standards require that male and female high school graduates achieve a minimum AFQT percentile score of 21; in addition, they are required to attain a percentile score of at least 30 on the General composite and a combined composite score (including the Mechanical, Administrative, General, and Electronics composites) of no less than 120.

Higher aptitude scores are required for male non-high school graduates and GED recipients in each of the Services. Female non-high school graduates are not eligible for enlistment in either the Navy or the Marine Corps; and female high school graduates who wish to enlist in these Services are required to meet different aptitude standards than those established for males.

The ASVAB scores of the profile study sample will be analyzed to estimate the numbers and porportions of American youth, by selected demographic variables, who would qualify for military enlistment under 1981 aptitude standards. The demographic categories selected for analysis are sex, racial/ethnic group, education, and geographic region.

Table 2

1981 Service Enlistment Aptitude Standards

(Required Operational Score on ASVAB 8 - 10)

Service/Education	Males		Females	
	Operational Standards		Operational Standards	
	AFQT	Aptitude Composites	AFQT	Aptitude Composites
Army				
High School Diploma Graduate	16	85 on 1	16	85 on 1
Non-High School Graduate (Including GED)	31	85 on 2	31	85 on 2
Navy				
High School Diploma Graduate	17	-		School Eligible[a]
GED	31	-		School Eligible[a]
Non-High School Graduate	38	-		Not Eligible
Marine Corps				
High School Diploma Graduate	21	GT[b] = 80	50	-
Non-High School Graduate (Including GED)	21	GT[b] = 95		Not Eligible
Air Force				
High School Diploma Graduate	21	G[c] = 30; MAGE[d] = 120	21	G[c] = 30; MAGE[d] = 120
GED	50	G[c] = 30; MAGE[d] = 120	50	G[c] = 30; MAGE[d] = 120
Non-High School Graduate	65	G[c] = 45; MAGE[d] = 170	65	G[c] = 45; MAGE[d] = 120

[a]Department of the Navy, "Criteria for selection of recruits and new accessions for formal school training." NAVMILPERSCOM Instruction 1236.1A. Washington, D.C.: Naval Military Personnel Command, Jan. 1981.

[b]General-Technical Composite

[c]General Composite

[d]Mechanical, Administrative, General Electronics Composites

THE PROFILE OF AMERICAN YOUTH marks the first time that a vocational aptitude battery has been given to a national probability sample. Up to this time, such research has not been conducted due to the great difficulty and expense involved in obtaining data. The subgroup analyses of current youth aptitudes and the projected qualification rates will soon be completed and released by the Department of Defense.

REFERENCES

Anastasi, A. Differential Psychology: Individual and Group Differences in Behavior (3rd Edition). New York: The MacMillan Company, 1958.

Bock, R. D. and Moore, E. G. J. Advantage and Disadvantage: Vocational Prospects of American Young People. Chicago, IL: National Opinion Research Center, October 1981.

Eitelberg, M. J. Subpopulation Differences In Performance on Tests of Mental Ability: Historical Review and Annotated Bibliography. Technical Memorandum 81-3. Washington, D. C.: Directorate for Assession Policy, Office of the Secretary of Defense, August 1981.

Featherman, D. L. "Schooling and Occupational Careers: Constancy and Change in World by Success." Constancy and Change in Human Development. Edited by G. Brian and J. Kagan. Cambridge, MA: Harvard University Press, 1980.

Jensen, A. R. Bias in Mental Testing. New York: The Free Press, 1980.

Maccoby, E. E. and Jacklin, C. M. The Psychology of Sex Differences. Stanford, CA: Stanford University Press, 1974.

Mathews, J. J., Valentine, L. D., and Sellman, W. S. Prediction of Reading Grade Levels of Service Applicants from Armed Services Vocational Aptitude Battery (ASVAB). AFHRL-TR-78-82. Brooks AFB, TX: Air Force Human Resources Laboratory, December 1978.

Scarr, S. Race, Social Class, and Individual Differences in I.Q. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1981.

Sewell, W. H. and Hauser, R. M. Education, Occupation, and Earnings. New York: Academic Press, 1975.

Tyler, L. E. The Psychology of Human Differences (3rd Edition). New York:

LEGAL AND POLITICAL CONSIDERATIONS IN LARGE-SCALE ADAPTIVE TESTING

Brian K. Waters

Gus C. Lee

As the term implies, adaptative testing is defined as a method of test construction wherein the items presented to a subject are selected iteratively dependent upon previous responses, thus "adapting" the test to the subject. Theoretically, such individual testing should provide more accurate measurement than group testing. Both simulated and live data studies have reported that the proportion of items required to reach a given level of reliability in a computer administered adaptive test (CAT) are about one-fourth to one-half of those required by a conventional paper-and-pencil test. Such dramatic efficiencies occur because after each item response, the computer program selects the next test item from the item pool which will provide the maximum amount of information about the examinee. McBride (1979) provides an excellent review of the advantages and possible disadvantages of CAT. There is little doubt that the use of interactive computer testing will increase enormously in the coming decade.

CAT is a technology preparing to make the transition from the laboratory to an operational environment. The vast majority of research and development in CAT has been sponsored by the military services, particularly the Navy, since Lord's early work on item response theory and CAT during the 1960s. Today, the Department of Defense (DoD) is sponsoring a large-scale, multi-year project to develop a CAT system for implementation in Armed Forces Examining and Entrance Stations across the country.

DoD assigned the Department of Navy responsibility for CAT development, with the Marine Corps as the executive agent. The Naval Personnel Research and Development Center (NPRDC) is currently in the process of selecting contractors to design, develop, and try out a prototype CAT system for delivering the Armed Services Vocational Aptitude Battery (ASVAB) adaptively to military applicants at nearly 1000 testing sites. Also, our host, ARI, has just released a "request-for-proposal" for a seven- year Army selection and classification system. CAT would likely be a part of such a system. Outside of DoD, the Coast Guard is sponsoring field research into CAT, and the commercial testing industry is investigating the large-scale use of CAT. Although no commerical tests have actually begun using CAT delivery, the American College Testing Program (ACT) and Educational Testing Service (ETS) are working toward that goal. Carol Dwyer's paper given yesterday afternoon at this conference mentioned that ETS is currently considering development of two-level sequential tests for a major admissions program.

Thus, the time has arrived when we must start considering some of the "real world" issues which CAT will face as actual decisions about examinees are made with adaptive testing. One such set of issues involves the legal and political considerations which could arise as CAT becomes an operational reality. This paper will discuss such legal and political issues. The authors will pose question--not answers--that need to be seriously considered as large-scale CAT approaches implementation.

To date very little, if anything, has been published on legal and political considerations of CAT. Warm (1978, p. 122) questioned the legal defensibility of having examinees take different numbers of items, or different sets of items. Wiskoff (1979) and Waters (1979) called for CAT researcher attention to such legal issues surrounding CAT.

CAT is a subset of testing in general. And, as I'm sure you all are well aware, testing has come under extreme pressure in the political and legal environments during the past decade. Strong political lobbies (Nader, 1979) have directed stinging criticism of commercial testing programs, and testing has become a frequent subject in litigation.

An annotated bibliography of court cases relevant to employment decisions (Cascio & Bernadin, 1981) was published by the Air Force Human Resources Laboratory (AFHRL). This useful document reviews 232 court cases from January 1971 - January 1980 dealing with adverse impact, unequal opportunity or pay, and bias in personnel selection, classification and evaluation systems. Each annotation provides the case reference, case source, court decision, critical cases cited as a basis for the decision, evidence of adverse impact, evidence of job-relatedness or validity, type of selection procedure, factors impacting the decision, effects of expert testimony, and implications for personnel policy. Not surprisingly, the authors of this paper found that at least 60 of these cases directly involved testing as a central issue. Overwhelmingly, the major focus of legal attention in these 60 cases was test validity, and the clear conclusion to be drawn was that job-related empirical validity was accepted unless practical constraints made empirical validation studies unrealistic. When empirical validation was impossible for a given test application, then content validation based upon careful job/task analysis was credible. Third in the list of court accepted validation methods was construct validity. Finally, face validity was given virtually no weight in the reviewed cases.

In a computerized adaptive testing mode, the validity issue portends possible problems in court. In CAT, it is not practical to validate the "test", only the item pool from which each test is drawn. This "validation" is completely different from empirical validations that the courts have previously accepted. Will this kind of predictive validity be satisfactory in a legal battle? Will the courts accept content validity when a very large number of items are said to measure a single trait? Rock & Bejar (1981) suggest that construct valdity is more defensible in CAT, but the courts have been hesitant to accept (or likely to understand) construct validation under conventional testing.

The latter point, that judges and juries may not understand many of the complicated technicalities involved with CAT, promises to be a major hurdle to CAT implementation. How do you explain latent trait theory, esoteric item selection strategies, "occult scoring methods" (Lord, 1978), and myriad other inexplicable jargon surrounding CAT to a court? Will a jurist accept the expert witness psychometrician who testifies that person A's theta is higher than person B's even though they have taken on items

in common; that person B answered a higher percentage of the items correctly; and that person B had to take twice as many items as person A to estimate his theta as accurately? Clearly the CAT community has a major educational and public relations chore in the court room and through the media before CAT will be legally accepted as a valid measurement procedure.

One example of such a problem in the political arena occurred with the ASVAB, where test scores are reported to Congress. A serious calibration error occurred on ASVAB in 1976, and the psychometrician's credibiltiy with the Congress suffered. It has been said that Congress feels that DoD phychometricians are some kind of amateur magicians out to perform statistical legerdemain. Imagine how they will react to latent trait theory?

The credibilty of CAT test scores may also be weakened by calibration problems like Douglass (1981) discussed. Parameter estimation methods and item linking procedures (Dorans, 1981) similarly are statistically complex activities which would be very difficult to explain in court or in Congress. We need to translate the scientific jargon that has become part of the CAT vocabulary into clear, concise, and comprehensible language that will communicate to the testing layman.

Another related issue is the recent truth-in-testing movement. After an initial strongly negative reaction by the testing industry to legislative directives on the release of test items to examinees, the major commercial test companies seem to have reluctantly accepted the concept. Computerized adaptive testing offers the prospect of making truth-in-testing more palatable to test developers. The release of items from a large pool to an examinee would not likely damage future administrations due to test compromise since theoretically every examinee takes an individually tailored test. Under current test development procedures, the new legislation will likely lead to expanded requirements for test items, additional validation studies, and increased test fees to examinees.

Theoretically, CAT and item response theory should reduce cultural test bias (Pine, et. al., 1979). The Cascio and Bernadin (1981) review cited many court cases in which alleged cultural test bias was a major issue, and numerous symposia, addresses, and paper sessions were presented at the APA convention two months ago on the same subject. Certainly it would be very beneficial to CAT's credibility in the courts if clear, unequivocal evidence should evolve which showed reduced cultural test bias under adaptive testing. At this time, this remains a research question.

This paper has strived to do no more than simply suggest questions that need to be considered as CAT nears operational usage. One thing we can be absolutley sure of is that once personnel selection and classification decisions begin to be made using CAT, there will be legal challenges to the validity of the measurement process. We need to anticipate these challenges, to conduct the research to answer the legal questions, and to understand enough about legal processes and judgments to "sell" the benefits of CAT to the courts and the public.

REFERENCES

Cascio, W. F., and Bernardin, H. J. Court Cases Relevant to Employment Decisions: Annotated Bibliography, ARHRL-TR-80-44. Brooks AFB, TX: Air Force Human Resources Laboratory, February 1981.

Dorans, N. J. Why and How to Insure that Items in the Same Pool are Appropriately Credentialed: Data Collection Strategies for Item Linking. Paper presented at the 23rd Military Testing Association Conference, Arlington, VA: October 1981.

Douglass, J. B., and McColskey, W. "A Comparison of Item Response Theory Item Calibration Techniques." Paper presented at the 23rd Military Testing Association Conference, Arlington, VA: October 1981.

McBride, J. R. Adaptive Mental Testing: The State of the Art. Technical Report 423. Washington, D.C.: US Army Research Institute for the Behavioral and Social Science, November 1079.

Nader, R. The Reign of ETS: The Corporation that Makes Up Minds. P. O. Box 19312, Washington, D.C. 20036

Pine, S. M., Church, A. T., Gialluca, K. A., and Weiss, D. J. Effects of Computerized Adaptive Testing on Black and White Students. Research Report 79-2, Minneapolis: University of Minnesota, March 1979.

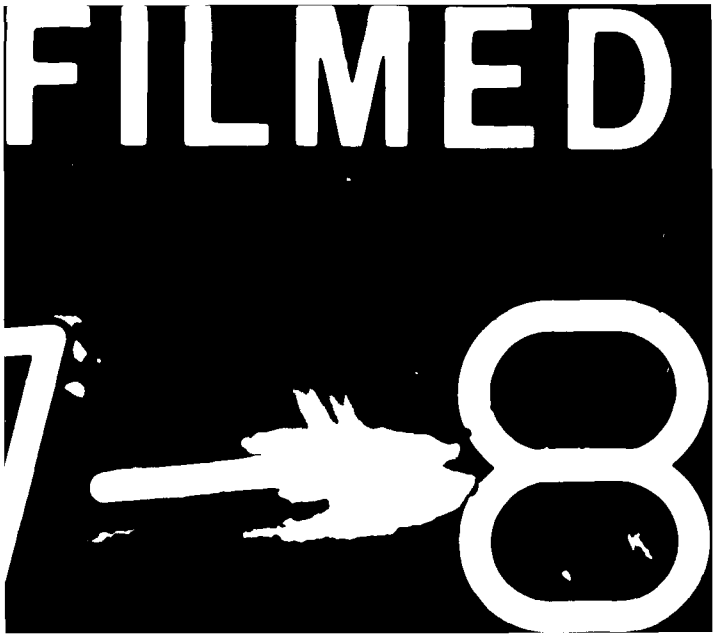
Pine, S. M., and Weiss, D. J. Bias-Free Computerized Testing: Final Report. Minneapolis: University of Minnesota, March 1979.

Rock, D. A., and Bejar, I. "Validity Considerations for Adaptive Testing Systems". Paper presented to the 23rd Annual Military Testing Association Conference, Arlington, VA: October 1981.

Warm, T. A. A Primer of Item Response Theory. Technical Report US-CG-941278, Oklahoma City, OK: U. S. Coast Guard Institute, 1978.

Waters, B. K. Discussant Remarks: Computerized Adaptive Testing Symposium. American Psychological Association Annual Meeting, New York: September 1979.

Wiskoff, M. Symposium Chair Remarks: Computerized Adaptive Testing Symposium. American Psychological Association Annual Meeting, New York: September 1979.


FILMED
7
8